PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-331490
(43)Date of publication of application : 30.11.2001

(51)Int.Cl. G06F 17/30
G06F 12/00
G06F 17/21

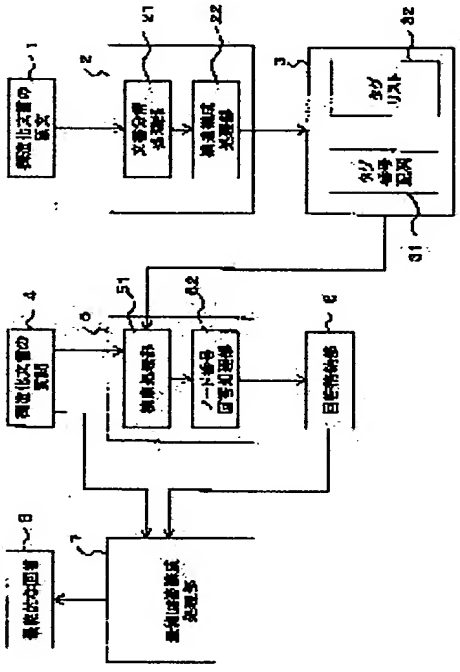
(21)Application number : 2001-062753
(22)Date of filing : 07.03.2001

(71)Applicant : FUJITSU LTD
(72)Inventor : NITTA KIYOSHI
KOZAKURA FUMIHIKO

(30)Priority
Priority number : 2000075678 Priority date : 17.03.2000 Priority country : JP

(54) STRUCTURED DOCUMENT STORAGE DEVICE, STRUCTURED DOCUMENT RETRIEVAL DEVICE, STRUCTURED DOCUMENT STORAGE AND RETRIEVAL DEVICE AND PROGRAM AND PROGRAM RECORDING MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To store a structured document suitable for retrieval at a high speed and the improvement of the using efficiency of memory resources regarding a structured document storage device.
SOLUTION: This structured document storage device is provided with a document decomposition processing part 21 for decomposing the original text 1 of the structured document into a tree structure composed of nodes and elements, a structure constitution processing part 22 for constituting a tag list structure based on the result of decomposition and a structure storage part 3 for storing at least the tag list structure. The tag list structure is composed of a tag number array 31 and a tag list 32. The tag number array 31 stores tag numbers uniquely determined for the respective kinds of tags which are information for indicating the kinds of the elements for the tags of the elements included in the original text 1 of the structured document and pointers to the tag list 32 corresponding to them. The tag list 32 is composed by connecting the nodes provided with the tags of the same kind or the tags of the different kinds in mutually same relation.



LEGAL STATUS
[Date of request for examination]
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－331490
（P2001－331490A）
(43)公開日　平成13年11月30日(2001.11.30)

(51)Int.Cl.⁷		識別記号	FI		テーマコード(参考)
Ｇ０６Ｆ	17/30	１４０	Ｇ０６Ｆ	17/30	１４０
		１７０			１７０Ａ
		２３０			２３０Ｚ
	12/00	５１３		12/00	５１３Ｚ
	17/21	５０１		17/21	５０１Ｔ

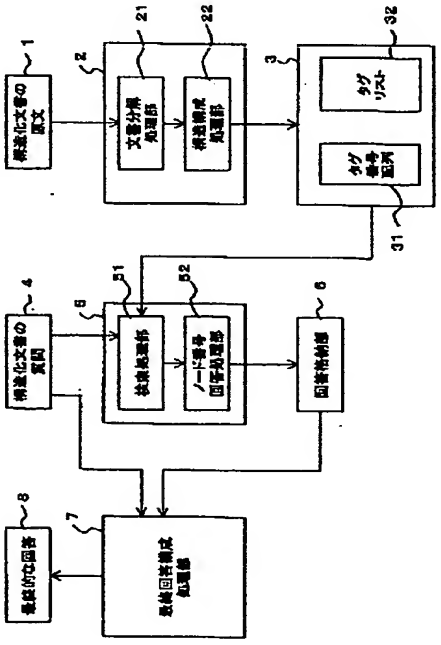
審査請求　未請求　請求項の数16　ＯＬ　（全 17 頁）

(21)出願番号　特願2001－62753(P2001－62753)
(22)出願日　平成13年３月７日(2001.3.7)
(31)優先権主張番号　特願2000－75678(P2000－75678)
(32)優先日　平成12年３月17日(2000.3.17)
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(72)発明者　新田　清
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　小櫻　文彦
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(74)代理人　100074848
弁理士　森田　寛　（外１名）

(54)【発明の名称】　構造化文書格納装置、構造化文書検索装置、構造化文書格納検索装置及びプログラム並びにプログラム記録媒体

(57)【要約】
【課題】本発明は、構造化文書格納装置に関し、高速での検索及びメモリ資源の使用効率の向上に適した構造化文書を格納することを目的とする。
【解決手段】　構造化文書格納装置は、構造化文書の原文１をノード及び要素からなる木構造に分解する文書分解処理部２１と、分解の結果に基づいて、タグリスト構造を構成する構造構成処理部２２と、少なくともタグリスト構造を格納する構造格納部３とを備える。タグリスト構造はタグ番号配列３１とタグリスト３２とからなる。タグ番号配列３１は、要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文１に含まれる要素のタグについてのタグ番号とこれに対応するタグリスト３２へのポインタとを格納してなる。タグリスト３２は、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つノードを連結してなる。



【特許請求の範囲】
【請求項１】　構造化文書の原文をノード及び要素からなる木構造に分解する文書分解処理部と、
前記分解の結果に基づいて、タグリスト構造を構成する構造構成処理部と、
少なくとも前記タグリスト構造を格納する構造格納部とを備え、
前記タグリスト構造はタグ番号配列とタグリストとからなり、
前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号とこれに対応する前記タグリストへのポインタとを格納してなり、
前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなることを特徴とする構造化文書格納装置。
【請求項２】　前記構造構成処理部が、当該構造化文書の原文から得られる前記ノード毎に、ノード番号を一意に定め、
前記タグリストにおいて、前記ノード番号を用いて前記ノードを連結することを特徴とする請求項１に記載の構造化文書格納装置。
【請求項３】　前記構造構成処理部が、当該構造化文書の原文から得られる前記要素毎に、要素番号を一意に定めることを特徴とする請求項２に記載の構造化文書格納装置。
【請求項４】　当該構造化文書格納装置が、更に、
前記構造化文書の原文を格納する原文格納部を備えることを特徴とする請求項１に記載の構造化文書格納装置。
【請求項５】　当該構造化文書格納装置が、更に、
前記構造化文書の原文を走査して、当該構造化文書の原文に含まれる要素についてのタグの種類の総数を算出するタグ算出処理部を備えることを特徴とする請求項１に記載の構造化文書格納装置。
【請求項６】　構造化文書の原文をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、前記構造化文書についての質問に基づいて検索する検索処理部を備え、
前記タグリスト構造はタグ番号配列とタグリストとからなり、
前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号及びこれに対応する前記タグリストへのポインタを格納してなり、
前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなり、
前記検索処理部が、前記質問の内容に該当するタグについてのタグ番号を用いて、前記タグ番号配列から当該タグ番号に対応する前記タグリストへのポインタを参照し、当該ポインタを用いて前記タグリストに連結された前記ノードを検索することを特徴とする構造化文書検索装置。
【請求項７】　前記タグリストにおいて、前記構造化文書の原文から得られる前記ノード毎に一意に定められるノード番号を用いて、前記ノードが連結され、
前記検索処理部が、当該検索の結果として、当該ノードのノード番号を得ることを特徴とする請求項６に記載の構造化文書検索装置。
【請求項８】　構造化文書の原文をノード及び要素に分解した木構造を用いて、前記構造化文書についての質問に対する回答を得るノード番号回答処理部を備え、
前記構造化文書の原文から得られる前記ノード毎にノード番号が一意に定められ、
前記ノード番号回答処理部が、前記質問に対する回答を、前記質問の内容に該当するノードのノード番号、又は、前記木構造において前記質問の内容に該当する範囲を規定するノードのノード番号で構成することを特徴とする構造化文書検索装置。
【請求項９】　当該構造化文書検索装置が、更に、
前記構造化文書の原文の文書名、前記質問の内容、及び、前記ノード番号で構成された組を、最終回答として構成する最終回答構成処理部を備えることを特徴とする請求項８に記載の構造化文書検索装置。
【請求項１０】　構造化文書の原文をノード及び要素からなる木構造に分解する文書分解処理部と、
前記分解の結果に基づいて、タグリスト構造を構成する構造構成処理部と、
少なくとも前記タグリスト構造を格納する構造格納部と、
構造化文書の原文をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、前記構造化文書についての質問に基づいて検索する検索処理部とを備え、
前記タグリスト構造はタグ番号配列とタグリストとからなり、
前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号とこれに対応する前記タグリストへのポインタとを格納してなり、
前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなり、
前記検索処理部が、前記質問の内容に該当するタグについてのタグ番号を用いて、前記タグ番号配列から当該タグ番号に対応する前記タグリストへのポインタを参照し、当該ポインタを用いて前記タグリストに連結された前記ノードを検索することを特徴とする構造化文書格納

検索装置。
【請求項１１】　コンピュータに実行させることにより構造化文書格納装置を実現するプログラムであって、
構造化文書の原文をノード及び要素からなる木構造に分解する文書分解処理と、
前記分解の結果に基づいて、タグリスト構造を構成する構造構成処理とを前記コンピュータに実行させ、
前記タグリスト構造はタグ番号配列とタグリストとからなり、
前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号とこれに対応する前記タグリストへのポインタとを格納してなり、
前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなることを特徴とするプログラム。
【請求項１２】　コンピュータに実行させることにより構造化文書格納装置を実現するプログラムを格納するコンピュータ読取可能なプログラム記録媒体であって、
前記プログラムは、
構造化文書の原文をノード及び要素からなる木構造に分解する文書分解処理と、
前記分解の結果に基づいて、タグリスト構造を構成する構造構成処理とを前記コンピュータに実行させ、
前記タグリスト構造はタグ番号配列とタグリストとからなり、
前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号とこれに対応する前記タグリストへのポインタとを格納してなり、
前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなることを特徴とするプログラム記録媒体。
【請求項１３】　コンピュータに実行させることにより構造化文書検索装置を実現するプログラムであって、
構造化文書の原文をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、前記構造化文書についての質問に基づいて検索する検索処理を前記コンピュータに実行させ、
前記タグリスト構造はタグ番号配列とタグリストとからなり、
前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号及びこれに対応する前記タグリストへのポインタを格納してなり、
前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなり、
前記検索処理が、前記質問の内容に該当するタグについてのタグ番号を用いて、前記タグ番号配列から当該タグ番号に対応する前記タグリストへのポインタを参照し、当該ポインタを用いて前記タグリストに連結された前記ノードを検索する処理であることを特徴とするプログラム。
【請求項１４】　コンピュータに実行させることにより構造化文書検索装置を実現するプログラムを格納するコンピュータ読取可能なプログラム記録媒体であって、
前記プログラムは、
構造化文書の原文をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、前記構造化文書についての質問に基づいて検索する検索処理を前記コンピュータに実行させ、
前記タグリスト構造はタグ番号配列とタグリストとからなり、
前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号及びこれに対応する前記タグリストへのポインタを格納してなり、
前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなり、
前記検索処理が、前記質問の内容に該当するタグについてのタグ番号を用いて、前記タグ番号配列から当該タグ番号に対応する前記タグリストへのポインタを参照し、当該ポインタを用いて前記タグリストに連結された前記ノードを検索する処理であることを特徴とするプログラム記録媒体。
【請求項１５】　コンピュータに実行させることにより構造化文書検索装置を実現するプログラムであって、
構造化文書の原文をノード及び要素に分解した木構造を用いて、前記構造化文書についての質問に対する回答を得るノード番号回答処理を前記コンピュータに実行させ、
前記構造化文書の原文から得られる前記ノード毎にノード番号が一意に定められ、
前記ノード番号回答処理が、前記質問に対する回答を、前記質問の内容に該当するノードのノード番号、又は、前記木構造において前記質問の内容に該当する範囲を規定するノードのノード番号で構成する処理であることを特徴とするプログラム。
【請求項１６】　コンピュータに実行させることにより構造化文書検索装置を実現するプログラムを格納するコンピュータ読取可能なプログラム記録媒体であって、
前記プログラムは、
構造化文書の原文をノード及び要素に分解した木構造を用いて、前記構造化文書についての質問に対する回答を

得るノード番号回答処理を前記コンピュータに実行させ、
前記構造化文書の原文から得られる前記ノード毎にノード番号が一意に定められ、
前記ノード番号回答処理が、前記質問に対する回答を、前記質問の内容に該当するノードのノード番号、又は、前記木構造において前記質問の内容に該当する範囲を規定するノードのノード番号で構成する処理であることを特徴とするプログラム記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、構造化文書格納装置、構造化文書検索装置、構造化文書格納検索装置及びプログラム並びにプログラム記録媒体に関し、特に、非定型な構造化文書を高速で効率よく検索することができる構造化文書格納装置、構造化文書検索装置、構造化文書格納検索装置及びプログラム並びにプログラム記録媒体に関する。

【０００２】

【従来の技術】企業等の組織においては、情報（データや知識）を当該組織内において流通、加工、共有することが求められる。このために、情報を共有することを目的とするシステム（ナレッジマネジメントシステム等）では、情報をＸＭＬ等のタグ付き言語で表現した文書（構造化文書）として作成して蓄積（格納）し、これを検索し、当該検索結果を利用する。このような構造化文書の利用は、多くの場合、以下の３つの方式のいずれかによっている。

【０００３】第１に、作成された構造化文書をそのまま格納し、その使用時に毎回、格納された構造化文書の全てを構造解析する方式がある。この方式は、例えば、特開平６－２５９４２１号公報「文書処理装置」等に開示されている。

【０００４】第２に、構造化文書を予め構造解析し、その結果を表形式のデータベースに変換して格納する方式がある。この方式は、例えば、特開平５－２２５２４０号公報「文書データベース装置」、特開平６－１１９３３１号公報「構造化文書の文書部品管理装置」、特開平７－４４５７９号公報「論理構造文書検索方式」、特開平７－３１９９１８号公報「文書検索対象指示装置」、特開平８－１４７３１１号公報「構造化文書検索方法及び装置」、特開平１０－１７１７９４号公報「動的部品化機能付き構造化文書データベースシステム」等に開示されている。

【０００５】第３に、構造化文書を予め構造解析し、その結果である「要素」と「構造情報」とを格納する方式がある。この方式は、例えば、特開平７－６５０３５号公報「構造化文書検索装置」、特開平１０－１８７６８０号公報「単語、文、部分の粒度で管理するドキュメントリポジトリ装置」等に開示されている。

【０００６】

【発明が解決しようとする課題】前述した第１乃至第３の方式には以下のような技術的な問題があることが、一般に知られている。

【０００７】即ち、第１の方式は、構造化文書の検索処理や回答（部分取り出し）処理等において格納された情報（構造化文書）を利用するが、各構造化文書をそのまま格納しているため、これについて利用の都度に構造解析を行なってから検索処理等を行なう。このため、検索処理等の効率が良くなく、処理速度も高速とは言えず、大規模なデータには適用が難しい。

【０００８】第２の方式は、データベース（Ｒ（関係）ＤＢ、表ＤＢ等）の構成（フィールドの構成）を予め決定し、構造化文書のどの部分を当該データベースのどのフィールドに格納するかを設計し、その後に当該データベースへの格納処理を行なう。従って、個々に（相互に）構造が異なるような複数の構造化文書の格納及び検索には適していない。また、通常のＲＤＢを用いる場合には、そのフィールドの最大サイズが決まっているので、解析結果に当該サイズ以上の部分を含む場合には、当該構造化文書のデータベースへの格納は難しい。

【０００９】第３の方式は、構造化文書の格納時にその構造解析を行なって、要素と要素間の構造情報に分離し、その状態のまま格納を行なう。第３の方式によれば、予め構造解析が行われた状態で格納されているので、第１の方式よりは検索処理等の効率がよく大規模なデータにも適用できる。また、構造解析の結果をそのまま格納しているので、格納（しようと）する構造化文書の構造やデータサイズの制約を受けないため、第２の方式よりは格納の自由度が高く、適用できる構造化文書の範囲が制限されにくい。

【００１０】しかし、第３の方式にも以下のような技術的な問題がある。第１に、第３の方式において格納される構造情報は、通常、木構造である。従って、検索処理において検索対象となる要素を取得するためには、「根（ルート）」と呼ばれる特定の要素から検索対象となる要素に至るまで、順に木構造をたどる必要がある。このため、検索速度が十分に高速であるとは言えない。また、木構造をたどる以外には検索対象となる要素を取り出す方法がなく、当該要素を直接取り出すことはできない。第２に、第３の方式において、構造情報は検索処理の手段にすぎず、検索処理の結果として得られるのは当該要素である。従って、検索処理の結果としては、当該結果である要素の実体を複製して蓄積しなければならず、当該結果の蓄積のためには多くのメモリ資源を消費する。このため、検索結果（複数の要素）を一時的に大量に蓄積して、これに対して（ソート等の）操作を行なってから、例えば上位１００件等を選別して出力しようとすると、膨大なメモリ資源を占有することになり、事実上困難である。

【００１１】本発明は、高速での検索及びメモリ資源の使用効率の向上に適した構造化文書を格納する構造化文書格納装置を提供することを目的とする。

【００１２】また、本発明は、構造化文書を高速で検索しその結果を効率よく格納する構造化文書検索装置を提供することを目的とする。

【００１３】また、本発明は、高速での検索及びメモリ資源の使用効率の向上に適した構造化文書を格納する構造化文書格納装置を実現するプログラム、及び、当該プログラムを記録するプログラム記録媒体を提供することを目的とする。

【００１４】また、本発明は、構造化文書を高速で検索しその結果を効率よく格納する構造化文書検索装置を実現するプログラム、及び、当該プログラム記録媒体を提供することを目的とする。

【００１５】

【課題を解決するための手段】図１は本発明の原理構成図であり、本発明による構造化文書格納検索システムの構成を示す。

【００１６】本発明の構造化文書格納装置は、構造化文書の原文１をノード及び要素からなる木構造に分解する文書分解処理部２１と、分解の結果に基づいて、タグリスト構造を構成する構造構成処理部２２と、少なくともタグリスト構造を格納する構造格納部３とを備える。タグリスト構造はタグ番号配列３１とタグリスト３２とからなる。タグ番号配列３１は、要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文１に含まれる要素のタグについてのタグ番号とこれに対応するタグリスト３２へのポインタとを格納してなる。タグリスト３２は、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つノードを連結してなる。

【００１７】本発明の構造化文書格納装置によれば、木構造に基づいて生成されたタグ番号配列３１とタグリスト３２とからなるタグリスト構造を得ることができる。このタグリスト構造は、後述するように、当該構造化文書についての質問４に基づく検索を高速に行うことができる構造である。従って、構造化文書の高速での検索に適したデータ構造を得ることができる。

【００１８】また、本発明の構造化文書検索装置は、構造化文書の原文１をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、構造化文書についての質問４に基づいて検索する検索処理部５１を備える。タグリスト構造は、前述と同様の構成である。検索処理部５１が、質問４の内容に該当するタグについてのタグ番号を用いて、タグ番号配列３１から当該タグ番号に対応するタグリスト３２へのポインタを参照し、当該ポインタを用いてタグリスト３２に連結されたノードを検索する。

【００１９】本発明の構造化文書検索装置によれば、木構造に基づいて生成されたタグ番号配列３１とタグリスト３２とからなるタグリスト構造を利用して検索を行う。即ち、タグ番号配列３１から構造化文書の原文１に含まれる要素のタグについてのタグ番号を求め、これを用いて対応するタグリスト３２へのポインタを求め、当該ポインタにより同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つノードを得る。従って、検索対象となる要素に至るまで順に木構造をたどることなく、かつ、検索対象となる要素以外の要素を全く検索することなく、検索対象となる要素を極めて高速に、検索対象となる要素を検索することができる。

【００２０】また、本発明の構造化文書検索装置は、構造化文書の原文１をノード及び要素に分解した木構造を用いて、構造化文書についての質問４に対する回答を得るノード番号回答処理部５２を備える。構造化文書の原文１から得られるノード毎にノード番号が一意に定められる。ノード番号回答処理部５２が、質問４に対する回答を、質問４の内容に該当するノードのノード番号、又は、木構造において質問４の内容に該当する範囲を規定するノードのノード番号で構成する。

【００２１】本発明の構造化文書検索装置によれば、質問４に対する回答が、一意に定められるノード番号により構成される。従って、検索処理の結果として当該結果である要素の実体を複製して蓄積する必要がなく、少ないメモリ資源で当該結果を蓄積することができる。このため、検索結果（複数の要素）を一時的に大量に蓄積して、これに対して（ソート等の）操作を行なって選別して出力するような処理を、容易に行うことができる。

【００２２】更に、本発明の一実施態様によれば、当該構造化文書検索装置が、更に、構造化文書の原文の文書名、質問の内容、及び、ノード番号で構成された回答の組を、最終回答として構成する最終回答構成処理部７を備える。

【００２３】これにより、質問４に対する回答が、短い文字列である文書名及び質問の内容と前述のノード番号とにより構成される。従って、検索処理の結果として当該結果である要素の実体を複製して蓄積する必要がなく、少ないメモリ資源で当該結果を蓄積することができる。

【００２４】以上の処理をコンピュータによって実現するためのプログラムは、コンピュータが読取可能な種々の半導体メモリ、ディスクメモリ、その他の可搬メモリ等の記録媒体に格納される。

【００２５】

【発明の実施の形態】図２は構造化文書格納検索システム構成図であり、本発明の構造化文書格納検索システムの構成を示す。構造化文書格納検索システムは、構造化文書格納装置である構造化文書格納処理部（処理モジュール、以下同じ）２、構造化文書検索装置の一部である構造化質問処理部５、構造化文書検索装置の一部である

最終回答構成処理部７からなる。これらの処理部は、１個のコンピュータ上で実現されてもよく、各々が独立したコンピュータ上で実現されてもよい。

【００２６】構造化文書格納処理部２は、文書分解処理部２１、構造構成処理部２２、原文格納処理部２５、構造格納部３、原文格納部９を備える。構造構成処理部２２は、ノード番号付与処理部２３、タグリスト構成処理部２４を備える。

【００２７】構造化文書格納処理部２は、与えられた（入力された）構造化文書の原文１を読み込む。原文格納処理部２５は、当該読み込まれた原文１（図３参照）を、必要に応じて、原文格納部９に格納する。従って、原文格納部９は、構造化文書の原文１をそのまま格納する。文書分解処理部２１は、当該読み込まれた構造化文書の原文１を分解処理して、当該原文１についてのノード及び要素からなる木構造３３（図４参照）を得る。木構造３３を得る手段は周知のいずれの手段によってもよい。

【００２８】構造構成処理部２２は、当該読み込まれた原文１の分解の結果である木構造３３に基づいて、ノード番号付与処理及びタグリスト構造を構成する処理を行う。即ち、ノード番号付与処理部２３が、木構造３３のノードにノード番号を付与し、要素に要素番号を付与する。このために、構造構成処理部２２又はノード番号付与処理部２３は、当該構造化文書の原文１から得られるノード毎に、ノード番号を一意に定め、タグリスト３２（図５参照）において、ノード番号を用いてノードを連結し、また、当該構造化文書の原文１から得られる要素毎に、要素番号を一意に定める。また、タグリスト構成処理部２４が、タグ番号配列３１（図５参照）とタグリスト３２とからなるタグリスト構造（図５参照）を構成する。木構造３３及びタグリスト構造は、構造格納部３に格納される。従って、構造格納部３は、検索処理のために、少なくともタグリスト構造を格納する。

【００２９】なお、構造化文書格納処理部２が、タグ算出処理部（図示せず）を備えるようにしてもよい。タグ算出処理部は、文書分解処理に先立って、前述の読み込まれた構造化文書の原文１を走査して、当該構造化文書の原文１に含まれる要素についてのタグの種類の総数を算出する。これにより、未知の構造化文書（例えば、文書型定義を参照することにより予め全てのタグを知ることができない文書）の原文１であっても、タグリスト構造の生成に先立って予めタグの総数を知り、後述するように、タグ番号を一意に定めることができる。

【００３０】構造化質問処理部５は、検索処理部５１、ノード番号回答処理部５２、回答格納部６を備える。構造化質問処理部５は、与えられた（入力された）構造化文書の質問４を読み込み、当該質問４に対する回答（一時回答）を作成して、回答格納部６に格納する。即ち、検索処理部５１が構造格納部３に格納されたタグリスト構造を用いて当該質問４に基づく検索を行い、この検索の結果を用いて、ノード番号回答処理部５２がノード番号を用いて表した回答（一時回答）を生成して、回答格納部６に格納する。

【００３１】検索処理部５１は、構造化文書の原文１をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、構造化文書についての質問４に基づいて検索する。検索処理部５１は、質問４の内容に該当するタグについてのタグ番号を用いて、タグ番号配列３１から当該タグ番号に対応するタグリスト３２へのポインタを参照し、当該ポインタを用いてタグリスト３２に連結されたノードを検索する。タグリスト３２において、構造化文書の原文１から得られるノード毎に一意に定められるノード番号を用いてノードが連結されているので、検索処理部５１は、当該検索の結果として、当該ノードのノード番号を得る。

【００３２】ノード番号回答処理部５２は、構造化文書の原文１をノード及び要素に分解した木構造を用いて、構造化文書についての質問４に対する回答を得る。ノード番号回答処理部５２は、質問４に対する回答を、質問４の内容に該当するノードのノード番号で構成する。又は、ノード番号回答処理部５２は、質問４に対する回答を、木構造において質問４の内容に該当する範囲を規定するノードのノード番号で構成する。いずれの処理を行うかは、質問４の内容に依存する。

【００３３】最終回答構成処理部７は、構造格納部３から木構造３３及びタグリスト構造を読み出し、回答格納部６から一時回答を読み出し、構造化文書の質問４を読み込み、これらに基づいて最終的な回答８を生成して出力する。具体的には、最終回答構成処理部７は、構造化文書の原文１の文書名、質問４の内容、及び、ノード番号で構成された組を、最終回答８として構成する。

【００３４】ここで、構造化文書の原文１、その木構造３３、そのタグリスト構造について説明する。

【００３５】図３は構造化文書の原文１の一例を示す。この文書の文書名（又はファイル名）は「０３２３．ｔｘｔ」であり、当該システムにおいて一意に定まり、これにより当該文書を識別できる。この原文１はＸＭＬで記述されているが、ＳＧＭＬ、ＨＴＭＬ等のタグ付き言語で記述されている文書であればよい。即ち、構造化文書はタグ付き言語で記述された文書（タグ付き文書）である。図３において、例えば、＜ｒｏｏｔ＞及び＜／ｒｏｏｔ＞等の「＜＞」で囲まれた部分がタグであり、「＜＞」及び「＜／＞」が一対をなす。また、例えば、タグ＜ｔｉｔｌｅ＞及び＜／ｔｉｔｌｅ＞に挟まれた文字列「ＡＡ社の株価」が要素である。

【００３６】図４は図３に示す構造化文書の原文１についての木構造３３を示す。即ち、当該原文１を要素に分解して得られる木構造３３を示す。木構造３３はノードと要素とを含む。ノード及び要素には、各々、ノード番

号及び要素番号が割り当てられる。図４において、角の丸い領域はタグに相当するノードを示す。ノードの横に記された数字はタグノードの通し番号を示す。ノード番号を用いた一時回答ではこの数字が用いられる。ノード番号には「＄」を付して表す。また、矩形領域はタグに囲まれた文書の要素を示す。これらの横に記された数字は要素の通し番号を示す。要素番号には「%」を付して表す。最上部の「ｒｏｏｔ」タグに相当するノードが、当該木構造３３のルートを示す。上下をつなぐ線分はノードどうしの親子関係（図中、上が親で下が子）を示す。なお、ノードの属性についても、ノード番号等と同様に、この木構造３３に融合して格納してもよい。

【００３７】従って、この明細書において、構造化文書の原文１を解析して分解した各要素を「ノード」、解析の結果構築する構造情報を「木構造」、ノードの種類を現わす情報を「タグ」（構造化文書表現におけるタグと同意）という。木構造３３の一部分であって、分岐を含まない部分木を「分岐なし部分木」という。即ち、「分岐なし部分木」とは、当該部分木の全ての構成要素について子へのリンク数が高々「１」であるものをいう。木構造３３における根（ルート）に相当するノードを「ルートノード」と呼ぶ。ルートからノードまでたどったときのタグのリストを「パス」と呼ぶ。構造化文書の重複しない名前又は識別子等を「文書のポインタ」と呼び、その他のデータについてもそれを識別できる情報を同様に「ポインタ」と呼ぶ。

【００３８】図５は図３に示す構造化文書の原文１についてのタグリスト構造を示す。タグリスト構造は、当該原文１及びこれを要素に分解して得られる木構造３３に基づいて得られる。タグリスト構造は、図５（Ａ）に示すように、タグ番号配列３１とタグリスト３２（３２Ａ）とからなる。

【００３９】タグ番号配列３１は、当該構造化文書の原文１に含まれる要素のタグについてのタグ番号と、これに対応するタグリスト３２へのポインタとを格納してなる。タグ番号は要素の種類を表す情報であるタグの種類毎に一意に定められる。具体的には、図５（Ｂ）に示すように、タグ番号配列３１は、タグとそのタグ番号とを格納する表である。例えば、要素「株価」の種類を表す情報であるタグ＜株価＞に対して、タグ番号「＃８０」が一意に定まる。タグ番号には＃を付して表す。タグ番号＃８０には、これに対応するタグリスト３２へのポインタＰ１が付加される。タグ番号はタグ番号配列３１のインデックスとなる。タグ番号インデックスに対応する文字列がタグ名である。

【００４０】タグ番号及び要素番号は、例えば次のように定められる。木構造３３の生成時において、各要素及びノードの原文１からの分離の順序をインデックスとし、ノードのポインタを値とする配列構造として生成する。当該配列におけるインデックス値が当該ノードのタグの通し番号となる。要素についても同様である。

【００４１】従って、図４及び図５（Ａ）に示すように、図３の文書１の先頭の２行（ＸＭＬでのきまりの部分）を無視して、次の＜ｒｏｏｔ＞から順にノード番号が与えられる。要素についても、最初に現われる「ＡＡ社の株価」から順に要素番号が与えられる。

【００４２】タグリスト３２は、同一種類のタグを持つノードを連結してなる。即ち、タグ＜株価＞を持つノードを連結している。図３及び図４において、タグ＜株価＞を持つノードはノード番号＄８のノード及びノード番号＄１１のノードであるので、ポインタＰ１及びＰ２により、これが連結される。タグ番号＃８０に付加されたポインタＰ１は、ノード番号＄８のノードを示す。ノード番号＄８のノードに付加されたポインタＰ２は、ノード番号＄１１のノードを示す。ノード番号＄１１のノードに付加されたポインタは、連結するノードがないので、空とされる。これらのポインタＰ１及びＰ２は、実際は、当該ノード自体ではなく、そのノード番号（の格納されている位置）を示す。

【００４３】なお、タグリスト３２は、同一種類のタグのみでなく、相互に同一の関係である異なる種類のタグを持つノードを連結してなるようにしてもよい。例えば、タグ番号＃７９であるタグ＜会社名＞は、＜企業名＞又は＜株式会社名＞等であっても同義である。従って、異なる種類のタグであっても、相互に同一の関係であるタグを持つノードを連結して、タグリスト３２を生成してもよい。

【００４４】構造化文書格納検索システムにおいては、以上の構成の構造化文書の原文１、木構造３３及びタグリスト構造（３１、３２）の格納処理と、これらを用いて構造化文書について質問４に基づく検索・回答処理が行われる。検索・回答処理としては、一時回答処理と最終回答処理とが行われる。

【００４５】格納処理においては、図６に示すように、格納する構造化文書の原文１を取得した上で、必要に応じて、複数の当該原文１Ａ乃至１Ｃの格納を行なう。原文１Ａの格納されたものを原文９２Ａと表す。原文１Ａと原文９２Ａとは同一である。複数の構造化文書の原文９２Ａ乃至９２Ｃが格納され、格納された文書群９１を構成する。文書分解処理により原文９２Ａ乃至９２Ｃの各々を、分解（タグで区切られる要素をそれぞれに分解）し、各々の木構造３３Ａ乃至３３Ｃを構成し、ノード番号及び要素番号を付与し、タグリスト構造を構成する。タグリスト構造において、タグ番号配列３１は文書群９１に対応して１個生成される。従って、タグ番号配列３１は原文９２Ａ乃至９２Ｃに共通であり、タグ番号は当該文書群９１において一意に定まる。一方、タグリスト３２Ａ乃至３２Ｃは、原文９２Ａ乃至９２Ｃの各々に対応して個々に生成される。

【００４６】なお、当該システムにおいて、文書群９１

を複数（図示せず）格納するようにしてもよい。この場合でも、図６に示すように、文書群９１毎にタグ番号配列３１が生成される。そして、１個の文書群９１及びタグ番号配列３１につき、これに含まれる文書１（９２）の数だけタグリスト３２が生成される。

【００４７】構造格納部３には、１個のタグ番号配列３１、タグリスト３２Ａ乃至３２Ｃ、木構造３３Ａ乃至３３Ｃ、ノード及び要素３４Ａ乃至３４Ｃが格納される。タグ番号配列３１、タグリスト３２Ａ乃至３２Ｃ、木構造３３Ａ乃至３３Ｃが全てノード番号、ポインタ等それ自体では意味のない情報からなるのに対して、ノード及び要素３４Ａ乃至３４Ｃは、各々、原文９２Ａ乃至９２Ｃの実際の要素及びノードそのものである。例えば、ノード及び要素３４Ａの前半部分には対応する原文９２Ａのノードが格納され、対応する木構造３３Ａの対応するノード番号によりポイントされ、後半部分には対応する原文９２Ａの要素が格納され、対応する木構造３３Ａの対応する要素番号によりポイントされる。

【００４８】検索・回答処理においては、図７に示すように、構造化文書の質問４を取得して、これに基づいて検索処理部５１が検索処理を行う。これにより当該質問の内容（条件）に該当する文書及びノード（ノード番号）を得る。この検索処理において、タグリスト構造（３１、３２）及び木構造３３を利用して、検索時における無駄な検索処理を抑制する。これにより、検索の高速化を図っている。

【００４９】即ち、どのような手法の構造検索を行なう場合でも、文書中に含まれる特定のタグ番号を持つ文書要素（ノード）を探す処理が行なわれる。タグリスト構造がない場合、その検索処理において、その都度ルートノードから文書全体をたどりつくす必要がある。これに対して、タグリスト構造は、その格納時に１回（タグ種類数が未知の場合は２回）木構造３３をたどることで生成できる。一旦生成してしまえば、検索時には文書全体をたどらずに、タグリスト構造にアクセスするだけで検索処理を行うことができる。これにより、ノード検索における実行ステップ数を削減し、検索時間を短縮することができる。なお、データの形態や検索条件によってその効果の程度は変化する。また、本発明のタグリスト構造による検索は、タグ番号からノードを検索する処理を短縮させるので、ノード数が多い文書が多数を占めるデータで、葉に近いノードを回答指示することが多い場合に比較的効果が大きい。

【００５０】ここで、構造化文書の質問４について説明する。図８（Ａ）は構造化文書の質問４の一例を示す。

【００５１】この質問４は、ＳＥＬＥＣＴ部、ＷＨＥＲＥ部、ＯＲＤＥＲＢＹ部からなる。この質問４は、タグ＜会社情報＞のすぐ内側にタグ＜会社名＞があり、タグ＜会社名＞で囲まれた内容が＜ＡＡ社＞であるような文書について（ＷＨＥＲＥ部）、条件を満たすタグ＜会社情報＞直下のタグ＜株価＞の内容を選択し（ＳＥＬＥＣＴ部）、その結果をその株価でソートせよ（ＯＲＤＥＲＢＹ部）、という指示を意味する。即ち、「ＡＡ社の株価を、その高い順に出力せよ」という回答指示である。なお、質問４において、ＳＥＬＥＣＴ部及びＯＲＤＥＲＢＹ部の「＃ｉ」は、ＳＥＬＥＣＴ部及びＯＲＤＥＲＢＹ部がＷＨＥＲＥ部の「＃ｉ」に連動するとを示す。即ち、ＳＥＬＥＣＴ部及びＯＲＤＥＲＢＹ部には＜ＡＡ社＞の記述がないが、＜ＡＡ社＞についての処理であることを指示している。

【００５２】ＳＥＬＥＣＴ部の内容「株価」が回答指示要素ｓである。回答指示要素は、ｓ１、ｓ２、・・・のように複数記述することができる。回答指示要素ｓ１、ｓ２、・・・が複数の場合、図８（Ｂ）に示すように、各々の要素毎に、回答（一時回答）の集合が得られる。回答の集合は、文書毎ではなく、要素毎に得られる。例えば、要素ｓ１について、回答の集合（ｄｐ，ｓ１，Ｎｎ（ｄ，ｓ））が得られる。各記号については後述する。

【００５３】図７において、このような質問４を取得した検索処理部５１は、ＳＥＬＥＣＴ部の「株価」を用いてタグ番号配列３１をアクセスして、図５（Ｂ）に示すように、そのタグ番号＃８０を知り、これを用いて文書群９１の全ての文書、即ち、全てのタグリスト３２を検索する。これにより、「株価」が文中に現れる文書及びそのノード（実際はノード番号及びそのポインタ）を、検索結果６０として得る。例えば、図３に示す文書の文書名「0323.txt」と、これにおける「株価」の現れるノード番号＄８及び＄１１を得る。この時点では、ノード番号＄１１はＢＢ社の株価であるので、回答としては採用できない。この検索結果６０は、例えば回答格納部６に一時的に格納される。従って、この検索において、関連のない文書及びノードを検索することなく、極めて高速に関連する文書名及びノード番号を得ることができる。

【００５４】このような検索結果６０を取得したノード番号回答処理部５２は、これに基づいて、第１ノード番号回答処理部５２１において質問４に対する回答６１を生成するか、又は、第２ノード番号回答処理部５２２において質問４に対する回答６２を生成する。回答６１は、質問４の内容に該当するノードのノード番号で構成する。例えば、ノード番号＄８である。回答６２は、木構造において質問４の内容に該当する範囲を規定するノードのノード番号で構成する。例えば、回答指示が＜ｒｏｏｔ＞＜ａｒｔｉｃｌｅ＞＜ｄａｔｅ＞である場合、上位のノード番号＄１及び下位のノード番号＄４である。

【００５５】このようにノード番号を用いた一時回答処理により回答格納部６に一時的な回答を格納し、最終回答処理により一時結果をソート等により選別し、ノード

番号の参照先は構造格納部３から得ることにより最終回答８を構成する。従って、ノード番号を格納し、検索回答にはそのノード番号を用いて一時的な結果を蓄積する。これにより、一時結果の蓄積時のメモリサイズを抑制することができる。一般に、構造検索におけるひとつの回答指定は原文１の一部（分岐なし部分木）となる。一時結果をノード番号を用いずに蓄積した場合、一時結果の件数及び文書サイズに比例する記憶領域が必要になる。しかし、一時結果をノード番号の形で蓄積することにより、メモリサイズを小さくでき、また文書サイズに影響を受けないようにできる。なお、一時回答を選別せずに全て出力する場合、ソート処理及び最終回答構成処理は省略してよい。

【００５６】その構成の手段が回答６１又は６２のいずれであっても、回答指示に対する直接の回答となるノードは、別に定まる。例えば、回答指示が「ＡＡ社の株価」であれば、前述の例において、ノード番号＄８のみが選択され、ＢＢ社の株価であるノード番号＄１１は選択されない。即ち、前述したように、ＷＨＥＲＥ部に従って、タグ＜会社情報＞のすぐ内側にタグ＜会社名＞があり、タグ＜会社名＞で囲まれた内容が＜ＡＡ社＞であるような文書について（従って、ノード番号＄１１は除かれる）、条件を満たすタグ＜会社情報＞直下のタグ＜株価＞の内容が選択される。この結果、例えば、図９（Ａ）に示すような結果が、回答６１又は６２として得られる。図９（Ａ）において、一行が一件の回答を示す。各回答は、ヒットした文書の文書名（を示すポインタ）ｄｐ、質問４の回答指示要素（ＳＥＬＥＣＴ部の指示）ｓ１、文書ｄｐ内のノード番号Ｎｎ（ｄ，ｓ）、ソート用の値からなる。

【００５７】回答６１又は６２は、操作処理部１０において、必要に応じて、ソート等の操作を受ける。なお、ソート等の処理は最終回答構成処理の後に行ってもよい。操作処理部１０は前述のＯＲＤＥＲＢＹ部を実行する。操作処理部１０は、図７に点線で示すように、当該構造化文書格納検索システム外のシステムに設けられていてもよい。即ち、当該処理はランキングスコア等によるソートのような周知の処理である。例えば、図９（Ａ）に示す回答６１又は６２をＯＲＤＥＲＢＹ部に従ってソートすると、図９（Ｂ）に示すような結果を得る。即ち、文書名「0323.txt」におけるノード番号＄８に対応する株価が最も高いという結果を得る。

【００５８】最終回答構成処理部７は、（ソート等された）回答６１を第１最終回答構成処理部７１において処理するか、又は、（ソート等された）回答６２を第２最終回答構成処理部７２において処理する。例えば、「ソートの結果の上位１件のみを出力せよ」という指示（図８（Ａ）には図示せず）であれば、図９（Ｃ）に示すように、上述のＡＡ社の最高の株価を出力する。

【００５９】図１０は、構造化文書格納検索システムが実行する構造化文書処理フローを示す。

【００６０】実行すべき処理が格納処理であるか否かを調べる（ステップＳ１）。

【００６１】格納処理である場合、格納処理を実行して（ステップＳ２）、処理を終了する。これについては、図１１を参照して後述する。

【００６２】格納処理でない場合、実行すべき処理が検索回答処理であるか否かを調べる（ステップＳ３）。

【００６３】検索回答処理である場合、検索回答処理を実行して（ステップＳ４）、処理を終了する。これについては、図１２を参照して後述する。

【００６４】検索回答処理でない場合、処理を終了する。

【００６５】図１１は、構造化文書格納処理部２が実行する構造化文書格納処理フローを示す。

【００６６】構造化文書格納処理部２が格納する構造化文書の原文１を取得する（ステップＳ１１）。

【００６７】構造化文書格納処理部２の原文格納処理部２５が、原文１を格納するか否かを調べる（ステップＳ１２）。格納しない場合、ステップＳ１４に進む。

【００６８】格納する場合、原文格納処理部２５が原文格納処理を行うことにより、当該原文１を原文格納部９に格納する（ステップＳ１３）。

【００６９】文書分解処理部２１が文書分解処理により当該原文１を分解し、構造構成処理部２２が構造構成処理によりノード番号配列、タグリスト３２、木を構成し、これらを構造格納部３に格納する（ステップＳ１４）。

【００７０】図１２は、構造化質問処理部５及び最終回答構成処理部７が実行する構造化質問処理及び回答処理フローを示す。

【００７１】検索処理部５１が、構造化文書の質問４を受け取る（ステップＳ２１）。

【００７２】検索処理部５１が、検索処理により該当文書を得る（ステップＳ２２）。これについては、図１３を参照して後述する。この図１３に示す処理は、各文書（１Ａ乃至１Ｃ）毎即ち、タグリスト３２（３２Ａ乃至３２Ｃ）毎に繰り返し行われる。これにより、質問４に該当する文書（の文書名）のみが得られる。

【００７３】ノード番号回答処理部５２が、ノード番号回答処理により回答格納部６に一時結果を格納する（ステップＳ２３）。これについては、図１４乃至図１６を参照して後述する。

【００７４】最終回答構成処理部７が、回答格納部６に格納された一時結果に基づいて、最終的な回答を構成して出力する（ステップＳ２４）。

【００７５】図１３は、検索処理部５１が実行する検索処理フローを示す。即ち、検索処理によりタグ番号からノードを取得する処理である。

【００７６】ノード返却集合Ｒを空にする（ステップＳ

３１）。

【００７７】タグリスト構造にタグ番号をキーにアクセスしタグリストポインタｐを取得する（ステップＳ３２）。

【００７８】タグリストポインタｐが未設定であるか否かを調べる（ステップＳ３３）。未設定である場合、ステップＳ３６に進む。

【００７９】未設定でない場合、当該ポインタｐの指すノードｎをノード返却集合Ｒに追加する（ステップＳ３４）。

【００８０】当該ポインタｐの次のタグリストポインタを、新しいタグリストポインタｐとして設定し（ステップＳ３５）、ステップＳ３３以下を繰り返す。

【００８１】以上により得たノード返却集合Ｒを、取得したノードとして返却し（ステップＳ３６）、処理を終了する。例えば、図５（Ｂ）に示すように、タグ＜株価＞からそのタグ番号を知り、これを用いてポインタＰ１及びＰ２の指すノード＄８及び＄１１をノード返却集合Ｒに得る。ノード返却集合Ｒが空である文書１（ｄ）は質問４に該当しない文書である。

【００８２】図１４は、ノード番号回答処理部５２及び最終回答構成処理部７が実行するノード番号回答処理及び最終回答構成処理フローを示す。

【００８３】ノード番号回答処理部５２が、検索質問４を満たす結果を全て抽出して、これを一時結果として回答格納部６に蓄積する（ステップＳ４１）。これについては、図１５を参照して後述する。

【００８４】最終回答構成処理部７が、必要に応じて、回答格納部６に蓄積された一時結果を、前述のように、ランキングスコアによるソート等の操作をする（ステップＳ４２）。この処理は、後述するように、必要に応じて実行される。

【００８５】最終回答構成処理部７が、操作の結果に基づいて、最終回答を生成する（ステップＳ４３）。これについては、図１６を参照して後述する。

【００８６】例えば、上位１０００件のランキング検索を行なう場合、ステップＳ４１で検索質問４を満たす結果を全て蓄積し、ステップＳ４２でランキングスコアによるソートを行ない、ステップＳ４３で上位１０００件について表示可能な回答を生成する。

【００８７】なお、ステップＳ４２は、必要に応じて、適宜実行される。即ち、上述の例において、ステップＳ４１において検索された結果である要素の個数が１０００件未満である場合、当該結果を全て出力できるので、ランキングスコアによるソートは必要ない場合もある。この場合、ステップＳ４２の実行は省略される。

【００８８】図１５は、ノード番号回答処理部５２が実行する一時結果の蓄積処理フローを示す。

【００８９】前処理を行う（ステップＳ５１）。即ち、検索を行ない、検索質問４にマッチする文書を文書集合Ｄにセットし、検索質問４の回答指示を要素毎に分解して回答指示要素集合Ｓにセットし、結果格納集合Ｒを空にする。

【００９０】文書集合Ｄが空か又は回答指示要素集合Ｓが空か否かを調べる（ステップＳ５２）。当該集合Ｄ又はＳが空の場合、ステップＳ６０へ進む。

【００９１】当該集合Ｄ及びＳが空でない場合、（ｄ，ｓ）∈（Ｄ，Ｓ）である（ｄ，ｓ）を１個取り出す（ステップＳ５３）。

【００９２】文書ｄ内で回答指示ｓにマッチする分岐なし部分木の最下部の（最も葉に近い）ノードの通し番号をＮｎ（ｄ，ｓ）にセットする（ステップＳ５４）。

【００９３】Ｎｎ（ｄ，ｓ）が空でないならば、文書ｄのポインタがｄｐ、回答指示要素ｓのポインタがｓｐのとき、ｒ（ｄｐ，ｓｐ，Ｎｎ（ｄ，ｓ））を結果格納集合Ｒに追加する（ステップＳ５５）。

【００９４】結果格納集合Ｒに結果ｒが複数個存在する場合は、１つにする（ステップＳ５６）。

【００９５】（ｄ，ｓ）∈（Ｄ，Ｓ）である全ての（ｄ，ｓ）について処理したか否かを調べる（ステップＳ５７）。処理していない場合、ステップＳ５３以下を繰り返す。

【００９６】処理した場合、結果格納集合Ｒが空か否かを調べる（ステップＳ５８）。空の場合、ステップＳ６０へ進む。

【００９７】空でない場合、回答指示要素集合Ｓと結果格納集合Ｒを一時結果として格納して、処理を終了する（ステップＳ５９）。

【００９８】結果格納集合Ｒを一時結果として格納する（ステップＳ６０）。例えば、回答指示要素ｓｐ＝＜株価＞について、図３乃至図５に示すｄｐ＝「0323.txt」なる文書のノード番号Ｎｎ（ｄ，ｓ）＝＄８を、結果格納集合Ｒに得る。即ち、（0323.txt，＜株価＞，＄８）を得る（図１９においても、略同様）。

【００９９】図１６及び図１７は、両者で１つのフローを構成し、最終回答構成処理部７が実行する最終回答構成処理フローを示す。

【０１００】前処理を行う（ステップＳ６１）。即ち、最終結果格納集合Ｒを空にする。ステップＳ４１（又はＳ４２）により格納した回答指示要素集合をＳｔ、一時結果格納集合をＲｔとする。

【０１０１】集合Ｓｔが空又は集合Ｒｔが空か否かを調べる（ステップＳ６２）。空の場合、ステップＳ６７へ進む。

【０１０２】空でない場合、ｒｔ（ｄｐ，ｓｐ，Ｎｎ（ｄ，ｓ））∈Ｒｔであるｒｔ（ｄｐ，ｓｐ，Ｎｎ（ｄ，ｓ））を１個取り出す（ステップＳ６３）。

【０１０３】文書ｄｐからノード番号配列（３４、図６参照）を用いてノード通し番号Ｎｎ（ｄ，ｓ）のノードｎを取得する（ステップＳ６４）。

【０１０４】回答指示要素集合Ｓｔからポインタｓｐの回答指示ｓを取得する（ステップＳ６５）。
【０１０５】木集合Ｔを空にする（ステップＳ６６）。
【０１０６】回答指示ｓが単一タグか否かを調べる（ステップＳ６７）。単一タグである（パス又はパスワイルドカードでない）場合、ステップＳ７１へ進む。ここで、単一タグとは、例えば図４において、ノード「株価」に要素「４，０２０」が繋がっているようなタグを言い、単一タグでないとは、例えば図４において、ノード「会社情報」にノード「会社名」及び要素「ＡＡ社」が繋がっているようなタグを言う。
【０１０７】単一タグでない（パス又はパスワイルドカードである）場合、回答指示ｓの分岐なし部分木の最上ノード（最もルートに近いノード）のタグ名ｔｎを取得する（ステップＳ６８）。
【０１０８】木構造情報を用いてノードｎからルートノードまで親方向へたどる（ステップＳ６９）。
【０１０９】当該たどる過程において、タグ名ｔｎと一致するノードｎｔがあったら、その各々について、ノードｎｔとノードｎの間の分岐なし部分木を、木集合Ｔに追加する（ステップＳ７０）。
【０１１０】ノードｎに子（子のノード）がある場合、木構造情報を用いてそれ以下の全てのノードを取得しノード集合Ｎｃとする（ステップＳ７１）。
【０１１１】ｔ∈Ｔである木ｔを１個取り出す（ステップＳ７２）。
【０１１２】ノード集合Ｎを空にする（ステップＳ７３）。
【０１１３】木構造ｔを構成するノードをノード集合Ｎに追加する（ステップＳ７４）。
【０１１４】全てのノード集合Ｎｃの要素をＮに追加する（ステップＳ７５）。
【０１１５】結果ｒ（ｄｐ，ｓｐ，Ｎ）を最終結果格納集合Ｒに追加する（ステップＳ７６）。
【０１１６】ｔ∈Ｔである全ての木ｔについて処理したか否かを調べる（ステップＳ７７）。全ての木ｔについて処理していない場合、ステップＳ７２以下を繰り返す。
【０１１７】全ての木ｔについて処理した場合、ｒｔ（ｄｐ，ｓｐ，Ｎｎ（ｄ，ｓ））∈Ｒｔである全てのｒｔ（ｄｐ，ｓｐ，Ｎｎ（ｄ，ｓ））について処理したか否かを調べる（ステップＳ７８）。全ての木ｔについて処理していない場合、ステップＳ６３以下を繰り返す。
【０１１８】全ての木ｔについて処理した場合、最終結果格納集合Ｒを結果とする（ステップＳ７９）。例えば、前述の（0323.txt，＜株価＞，＄８）に基づいて、文書0323.txtのノード＄８の内容「株価」をノード（配列）３４から得て、これにその要素％５「４，０２０」を追加してＮとし、結果ｒ（0323.txt，＜株価＞，株価及び４，０２０）を集合Ｒに得る（図２０においても、略同様）。
【０１１９】以上に述べた図１４に示すノード番号回答処理及び最終回答構成処理に代えて、以下の図１８に示すように、ノード番号回答処理及び最終回答構成処理を実行してもよい。図１８は、ノード番号回答処理部５２及び最終回答構成処理部７が実行するノード番号回答処理及び最終回答構成処理フローを示す。
【０１２０】ノード番号回答処理部５２が、ステップＳ４１と同様に、検索質問４を満たす結果を全て抽出して、これを一時結果として回答格納部６に蓄積する（ステップＳ８１）。これについては、図１９を参照して後述する。
【０１２１】最終回答構成処理部７が、ステップＳ４２と同様に、必要に応じて、回答格納部６に蓄積された一時結果を、前述のように、ランキングスコアによるソート等の操作をする（ステップＳ８２）。この処理は、前述のステップＳ４２と同様に、必要に応じて実行される。
【０１２２】最終回答構成処理部７が、ステップＳ４３と同様に、操作の結果に基づいて、最終回答を生成する（ステップＳ８３）。これについては、図２０を参照して後述する。
【０１２３】図１９は、ノード番号回答処理部５２が実行する一時結果の蓄積処理フローを示す。
【０１２４】前処理を行う（ステップＳ９１）。即ち、検索を行ない、検索質問４にマッチする文書を文書集合Ｄにセットし、検索質問４の回答指示を要素毎に分解して回答指示要素集合Ｓにセットし、結果格納集合Ｒを空にする。
【０１２５】文書集合Ｄが空か又は回答指示要素集合Ｓが空か否かを調べる（ステップＳ９２）。当該集合Ｄ又はＳが空の場合、ステップＳ９９へ進む。
【０１２６】当該集合Ｄ及びＳが空でない場合、（ｄ，ｓ）∈（Ｄ，Ｓ）である（ｄ，ｓ）を１個取り出す（ステップＳ９３）。
【０１２７】文書ｄ内で回答指示ｓにマッチする分岐なし部分木の最上部の（最もルートに近い）ノードの通し番号をＮｒ（ｄ，ｓ）にセットし、最下部の（最も葉に近い）ノードの通し番号をＮｌ（ｄ，ｓ）にセットする（ステップＳ９４）。
【０１２８】Ｎｒ（ｄ，ｓ）及びＮｌ（ｄ，ｓ）が空でないならば、文書ｄのポインタがｄｐ、回答指示要素ｓのポインタがｓｐのとき、ｒ（ｄｐ，ｓｐ，Ｎｒ（ｄ，ｓ），Ｎｌ（ｄ，ｓ））を結果格納集合Ｒに追加する（ステップＳ９５）。
【０１２９】（ｄ，ｓ）∈（Ｄ，Ｓ）である全ての（ｄ，ｓ）について処理したか否かを調べる（ステップＳ９６）。処理していない場合、ステップＳ９３以下を繰り返す。
【０１３０】処理した場合、結果格納集合Ｒが空か否か

を調べる（ステップＳ９７）。空の場合、ステップＳ９９へ進む。

【０１３１】空でない場合、回答指示要素集合Ｓと結果格納集合Ｒを一時結果として格納して、処理を終了する（ステップＳ９８）。

【０１３２】結果格納集合Ｒを一時結果として格納する（ステップＳ９９）。

【０１３３】図２０は、最終回答構成処理部７が実行する最終回答構成処理フローを示す。

【０１３４】前処理を行う（ステップＳ１０１）。即ち、最終結果格納集合Ｒを空にする。ステップＳ８１（又はＳ８２）により格納した回答指示要素集合をＳｔ、一時結果格納集合をＲｔとする。

【０１３５】集合Ｓｔが空又は集合Ｒｔが空か否かを調べる（ステップＳ１０２）。空の場合、ステップＳ１１０へ進む。

【０１３６】空でない場合、ｒｔ（ｄｐ，ｓｐ，Ｎｒ（ｄ，ｓ），Ｎｌ（ｄ，ｓ））∈Ｒｔであるｒｔ（ｄｐ，ｓｐ，Ｎｒ（ｄ，ｓ），Ｎｌ（ｄ，ｓ））を１個取り出す（ステップＳ１０３）。

【０１３７】ノード集合Ｎを空にする（ステップＳ１０４）。

【０１３８】文書ｄｐからノード番号配列を用いてノード通し番号Ｎｒ（ｄ，ｓ）のノードｎｒ、Ｎｌ（ｄ，ｓ）のノードｎｌを取得する（ステップＳ１０５）。

【０１３９】木構造情報を用いてノードｎｌからノードｎｒまでたどり、それらのノードを全てノード集合Ｎに追加する（ステップＳ１０６）。

【０１４０】ノードｎｌに子（子のノード）がある場合は木構造情報を用いてそれ以下の全てのノードを取得しノード集合Ｎに追加する（ステップＳ１０７）。

【０１４１】結果ｒ（ｄｐ，ｓｐ，Ｎ）を最終結果格納集合Ｒに追加する（ステップＳ１０８）。

【０１４２】ｒｔ（ｄｐ，ｓｐ，Ｎｒ（ｄ，ｓ），Ｎｌ（ｄ，ｓ））∈Ｒｔである全てのｒｔ（ｄｐ，ｓｐ，Ｎｒ（ｄ，ｓ），Ｎｌ（ｄ，ｓ））について処理したか否かを調べる（ステップＳ１０９）。処理していない場合、ステップＳ１０３以下を繰り返す。

【０１４３】処理した場合、最終結果格納集合Ｒを結果として終了する（ステップＳ１１０）。

【０１４４】なお、図１４乃至図１７と図１８乃至図２０との対比から判るように、これらの処理では、回答時の処理が少し異なる。回答処理の一時格納においては、前者は回答指示の葉の部分のみ蓄積するが、後者は回答指示のルートの部分も蓄積する。従って、一時蓄積する回答の分量は前者の方が少なくて済む。一方、回答処理の最終回答生成においては、前者は回答指示のルートに相当するノードを検索し直すが、後者はその処理を行なう必要がない。従って、最終回答生成に要する処理の時間は後者のほうが短かくて済む。

【０１４５】

【発明の効果】以上説明したように、本発明によれば、構造化文書格納装置において、構造化文書の原文に対応する木構造に基づいて生成されたタグ番号配列とタグリストとからなるタグリスト構造を備えることにより、構造化文書の高速での検索に適したデータ構造を得ることができる。

【０１４６】また、本発明によれば、構造化文書検索装置において、木構造に基づいて生成されたタグ番号配列とタグリストとからなるタグリスト構造を利用して検索を行うことにより、検索対象となる要素に至るまで順に木構造をたどることなく、かつ、検索対象となる要素以外の要素を全く検索することなく、検索対象となる要素を極めて高速に、検索対象となる要素を検索することができる。

【０１４７】また、本発明によれば、構造化文書検索装置において、質問に対する回答をノード番号により構成することにより、検索処理の結果として当該結果である要素の実体を複製して蓄積する必要がなく、少ないメモリ資源で当該結果を蓄積することができるので、検索結果を一時的に大量に蓄積して、これに対して操作を行なって選別して出力するような処理を、容易に行うことができる。

【図面の簡単な説明】

【図１】本発明の原理構成図である。

【図２】構造化文書格納検索システム構成図である。

【図３】原文説明図である。

【図４】木構造説明図である。

【図５】タグリスト構造説明図である。

【図６】構造化文書格納検索説明図である。

【図７】構造化文書格納検索説明図である。

【図８】構造化文書格納検索説明図である。

【図９】構造化文書格納検索説明図である。

【図１０】構造化文書処理フローを示す。

【図１１】構造化文書格納処理フローを示す。

【図１２】構造化質問処理及び回答処理フローを示す。

【図１３】検索処理フローを示す。

【図１４】ノード番号回答処理及び最終回答構成処理フローを示す。

【図１５】一時結果の蓄積処理フローを示す。

【図１６】最終回答構成処理フローを示す。

【図１７】最終回答構成処理フローを示す。

【図１８】ノード番号回答処理及び最終回答構成処理フローを示す。

【図１９】一時結果の蓄積処理フローを示す。

【図２０】最終回答構成処理フローを示す。

【符号の説明】

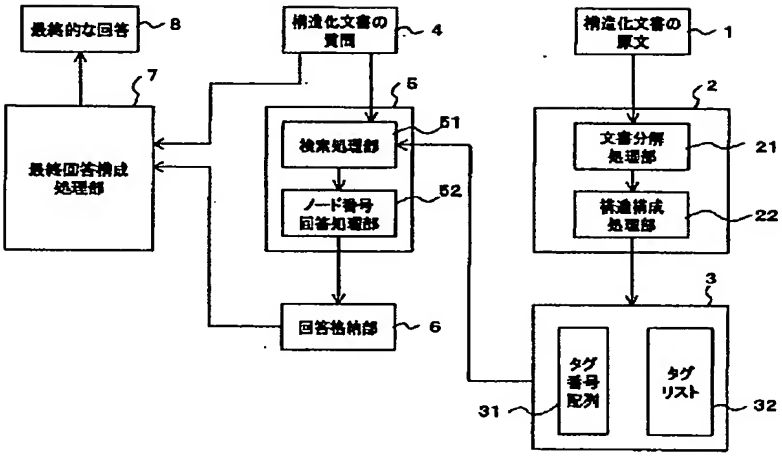
３　構造格納部

６　回答格納部

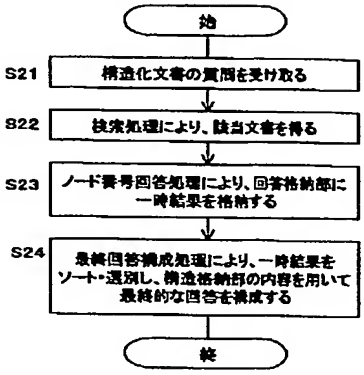
２１　文書分解処理部

２２　構造構成処理部
２３　ノード番号付与処理部
２４　タグリスト構成処理部
３１　タグ番号配列
３２　タグリスト構造（群）
５１　検索処理部
５２　ノード番号回答処理部
７１　第１最終回答処理部
７２　第２最終回答処理部

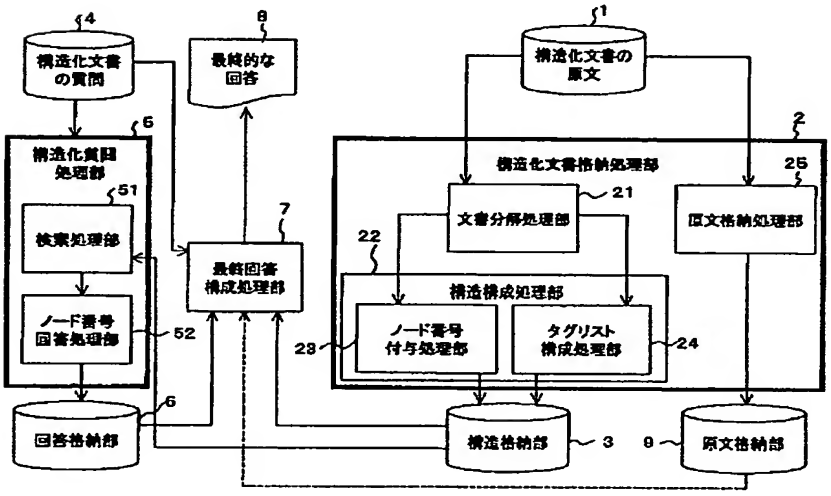
【図１】



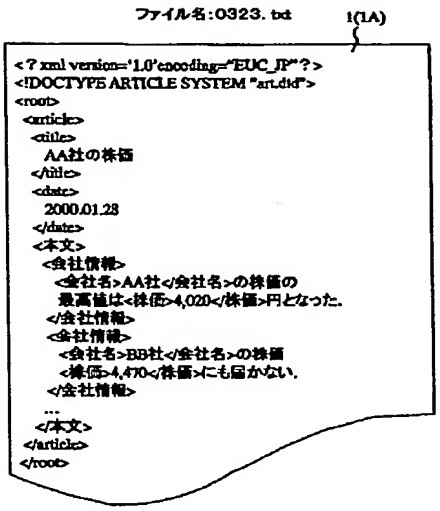
【図１２】



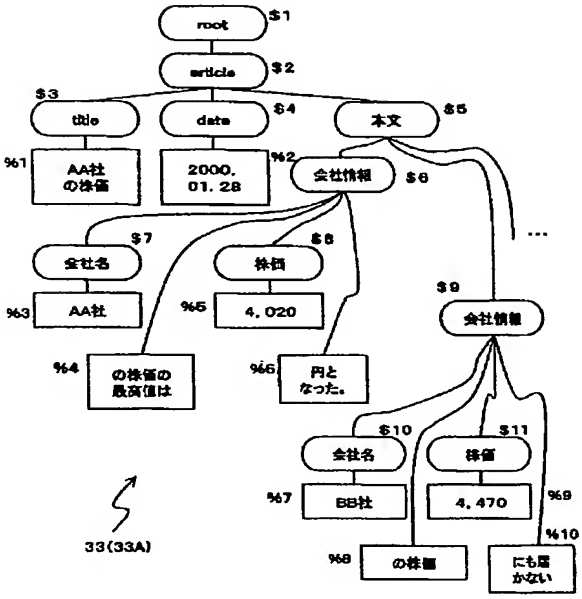
【図２】



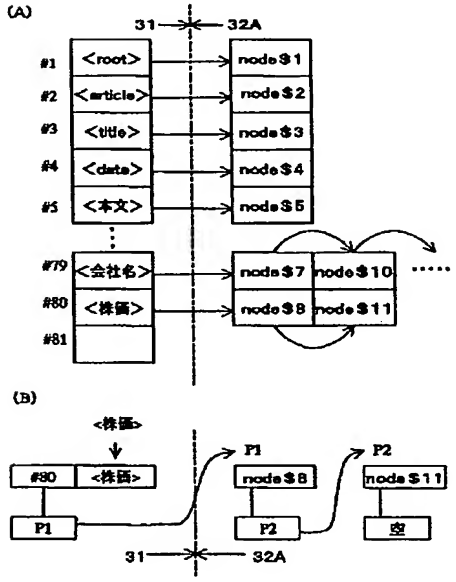
【図３】


ファイル名:0323. txt
1(1A)
<?xml version='1.0'encoding="EUC_JP"?>
<!DOCTYPE ARTICLE SYSTEM "art.dtd">
<root>
<article>
<title>
AA社の株価
</title>
<date>
2000.01.28
</date>
<本文>
<会社情報>
<会社名>AA社</会社名>の株価の
最高値は<株価>4,020</株価>円となった.
</会社情報>
<会社情報>
<会社名>BB社</会社名>の株価
<株価>4,470</株価>にも届かない.
</会社情報>
...
</本文>
</article>
</root>

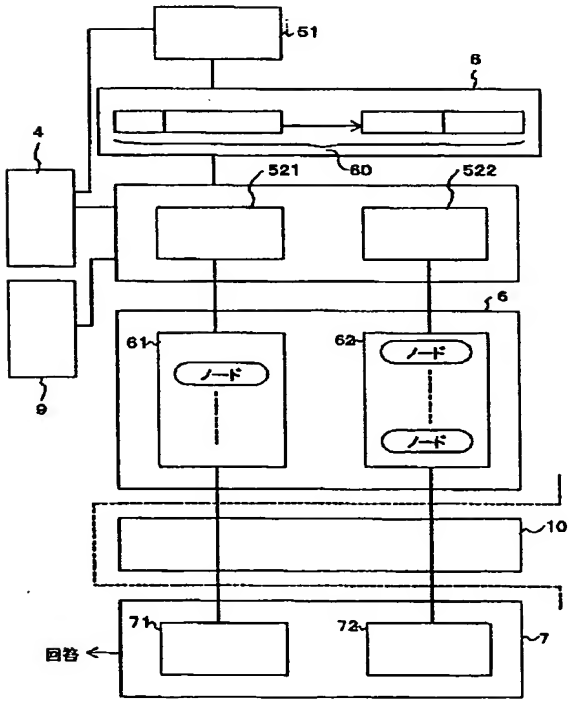
【図４】


root
$1
article
$2
$3
title
date
$4
本文
$5
%1
AA社
の株価
2000.
01. 28
%2
会社情報
$6
$7
会社名
株価
$8
%3
AA社
%5
4, 020
$9
会社情報
%4
の株価の
最高値は
%6
円と
なった。
$10
$11
会社名
株価
%7
BB社
4, 470
%9
%10
%8
の株価
にも届
かない
33(33A)

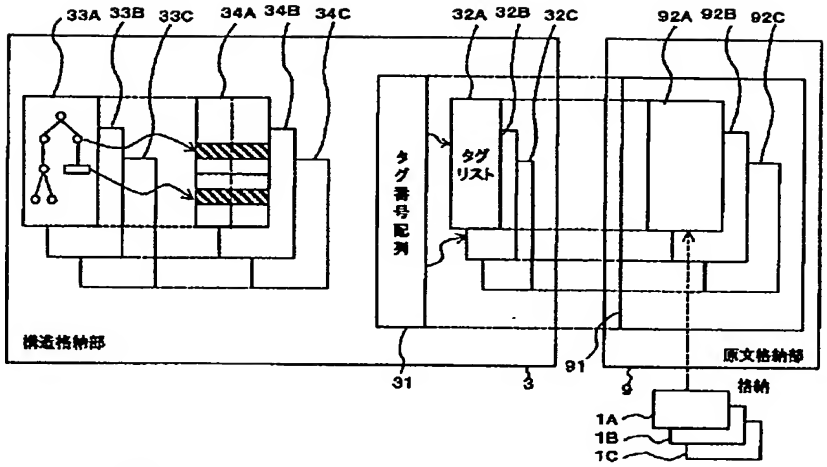
【図５】


(A)
31
32A
#1
<root>
node$1
#2
<article>
node$2
#3
<title>
node$3
#4
<date>
node$4
#5
<本文>
node$5
#79
<会社名>
node$7
node$10
#80
<株価>
node$8
node$11
#81
(B)
<株価>
#80
<株価>
P1
P1
node$8
P2
P2
node$11
空
31
32A

【図７】


51
6
4
521
60
522
6
61
62
ノード
ノード
ノード
9
10
71
72
7
回答

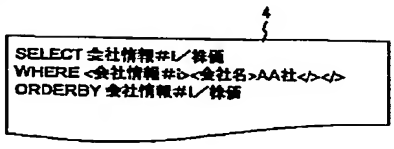
【図６】



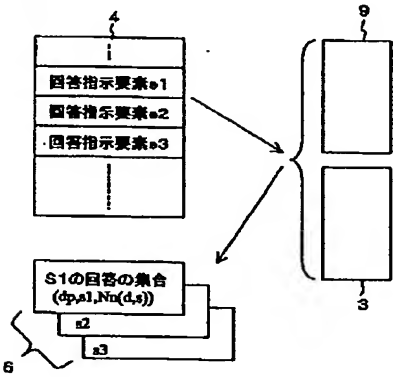
【図８】

【図９】

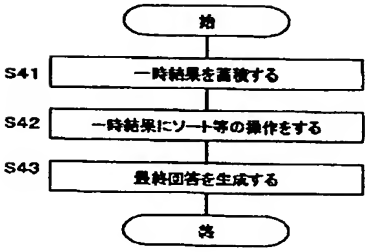
(A)



(B)



【図１４】



(A)

文書名	回答指示要素	ノード通し番号	ソート値
0001	株価	8	3, 670
0197	株価	8	3, 090
0224	株価	14	3, 820
0323	株価	8	4, 020
0457	株価	22	3, 800
0598	株価	57	2, 380

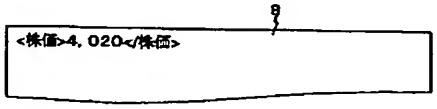
6

(B)

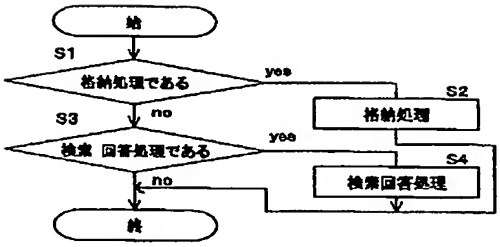
文書名	回答指示要素	ノード通し番号	ソート値
0323	株価	8	4, 020
0224	株価	14	3, 820
0457	株価	22	3, 800
0001	株価	8	3, 670
0197	株価	8	3, 090
0598	株価	57	2, 380

6

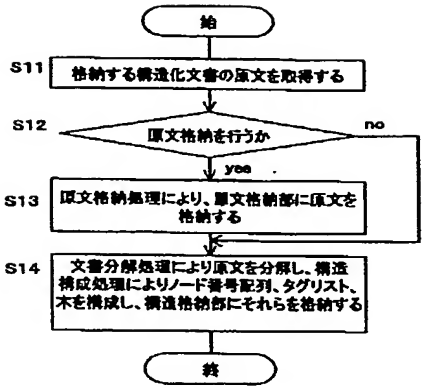
(C)



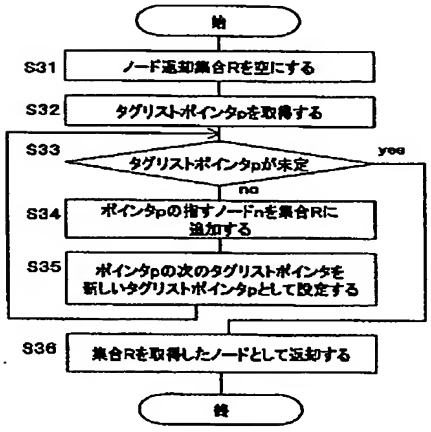
【図１０】



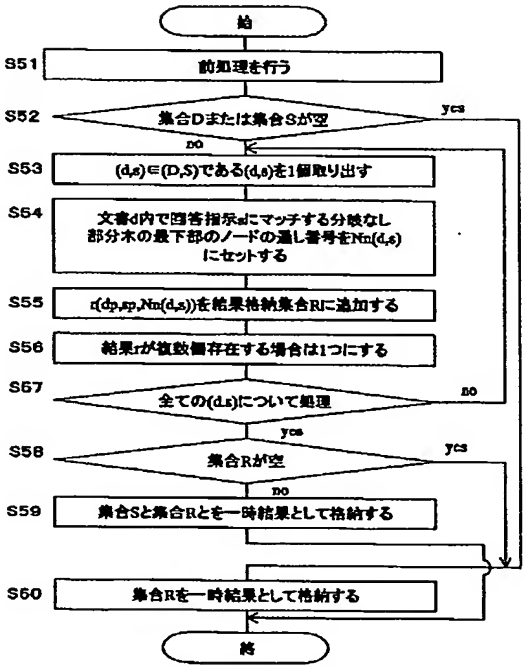
【図１１】



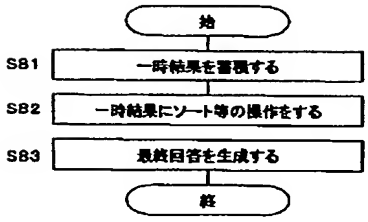
【図１３】



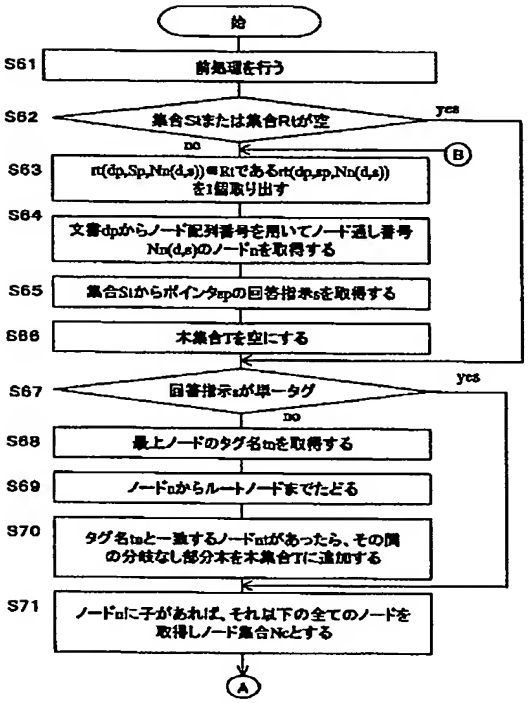
【図１５】



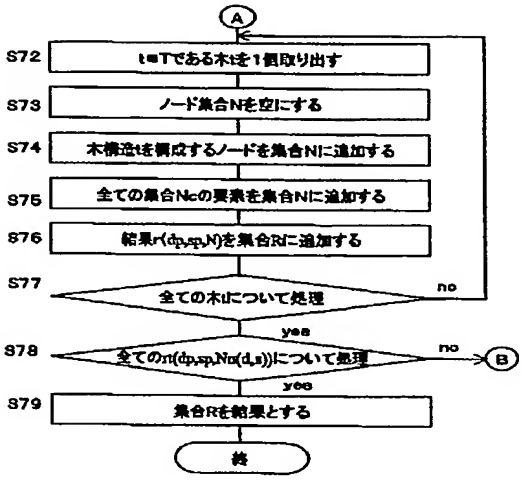
【図１８】



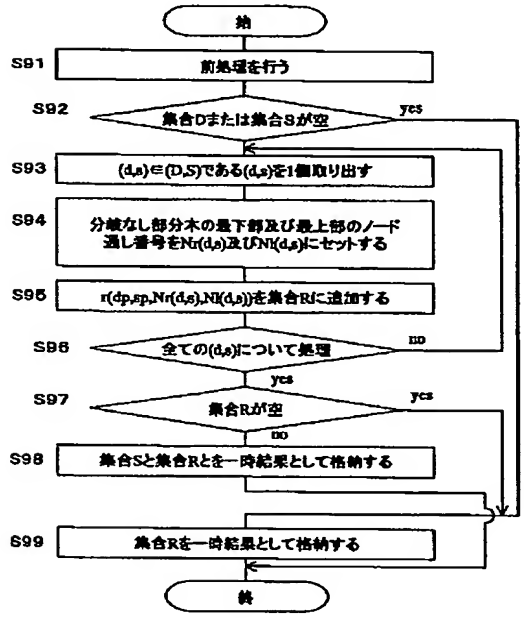
【図１６】



【図１７】



【図１９】



【図２０】